Analisi PCA

ELABORAZIONE DEI DATI SCIENTIFICI

[262-009]

Autore:

Testoni Manuel (219155)

November 13, 2025

Contents

1

Introduzione

Nel corso dell'analisi sono stati generati molteplici grafici quali: dendrogramma, ognuno con le varie combinazioni per distanza e tipo di linkage, grafici di clustering tramite K-Means, tramite PC's e grafici di clustering di variabili. Tramite questi grafici si è in grado di operare un'analisi di clustering sul dataset in questione andando a discriminare quali siano i metodi, parametri ed iper parametri migliori per un raggruppamento ottimale delle varie classi del dataset.

1.1 Obiettivo

Questo report mira a fare un'analisi di tipo clustering sul dataset Olive oil già visto in precedenza, che tratta quindi di campioni di olii di oliva italiani, andando a confrontare tecniche di clustering gerarchico con tecniche di clustering di tipo patitioning, quali K-Means.

1.2 Descrizione del Dataset

Il dataset *Olive Oil* contiene misure chimiche relative alla composizione in acidi grassi di 572 campioni di olio d'oliva provenienti da diverse regioni. Ogni osservazione rappresenta un campione di olio, mentre le variabili descrivono la concentrazione (espressa in unità relative) dei principali acidi grassi che ne determinano la composizione e la qualità.
Le variabili considerate sono:

- **Palmitic**: acido grasso saturo predominante, indicatore della stabilità ossidativa dell'olio;
- **Palmitoleic**: acido grasso monoinsaturo, associato a caratteristiche varietali;
- **Stearic**: acido saturo secondario, generalmente presente in quantità moderate;
- **Oleic**: acido grasso monoinsaturo principale, che rappresenta la frazione più abbondante dell'olio e ne influenza le proprietà nutrizionali;
- **Linoleic**: acido polinsaturo, legato alla freschezza e alla stabilità ossidativa;
- **Eicosanoic**: acido grasso a catena lunga, presente in piccole quantità;
- **Linolenic**: acido grasso polinsaturo essenziale, anch'esso presente in basse concentrazioni.

Table 1.1: Statistiche descrittive delle variabili del dataset *Olive Oil*.

Variabile	Media	Dev. Std.	Min	25%	Mediana	Max
Palmitic	1231.74	168.59	610	1095.00	1201.0	1753.0
Palmitoleic	126.09	52.49	15	87.75	110.0	280.0
Stearic	228.87	36.74	152	205.00	223.0	375.0
Oleic	7311.75	405.81	6300	7000.00	7302.5	8410.0
Linoleic	980.53	242.80	448	770.75	1030.0	1470.0
Eicosanoic	31.89	12.97	0	26.00	33.0	74.0
Linolenic	58.10	22.03	0	50.00	61.0	105.0

2

Risultati

2.1 Scelta configurazione Distanza-Linkage

I grafici generati sono state tutte le combinazioni tra i seguenti linkage methods:

- ward
- single
- complete
- average
- centroid

e le seguenti distanze:

- Euclidean
- Correlation
- Minkowski

Tra questi 12 grafici generati sono stati scelti i due con il silhouette score più elevato. Questo in quanto il silhouette score rappresenta quanto bene ogni punto è assegnato al proprio cluster. Questo significa che cluster con silhouette score più alto avranno una struttura più compatta, con maggiore separazione e non confusi; Per questo si tende a prediligere cluster con silhouette score più alto.

Le due combinazioni che hanno generato il silhouette score più alto sono state: **Complete Linkage + Minkowski** e **Average + Minkowski**.

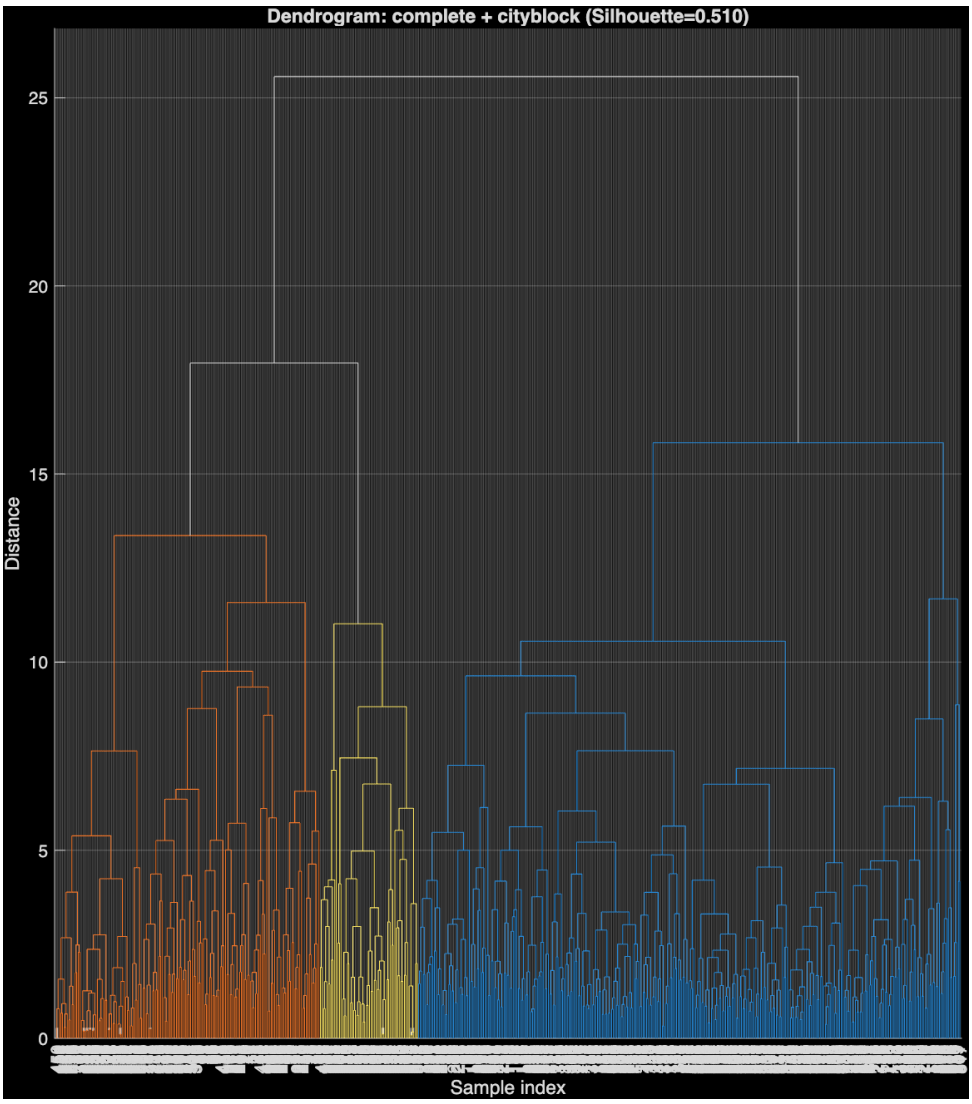


Figure 2.1: Dendrogramma del dataset Olive Oil tramite tecnica di clustering gerarchico complete.

Il dendrogramma in Figura **??** rappresenta la struttura gerarchica dei cluster ottenuti mediante il metodo di collegamento *complete* e la distanza *CityBlock* (o *Manhattan*), che non è altro che un caso particolare della distanza di minkowski quando il parametro "p" risulta essere uguale a 1. Il parametro p, è quel parametro che regola quanto peso dare alle differenze calcolate all'interno della formula. L'asse delle ascisse mostra i campioni, mentre l'asse delle ordinate indica la distanza alla quale i cluster vengono uniti. Da questo primo dendrogramma siamo in grado di notare che si creano 3 cluster: giallo, arancione e blu. I cluster arancio e giallo si uniscono a distanze ravvicinate quindi indicano una forte similarità. Al contrario il cluster blu si unisce nel ramo con gli altri due cluster ad una distanza elevatissima, questo suggerisce che vi sia una forte dissimilarità.

Inoltre il valore di silhouette è globalmente buono anche se non perfetto.

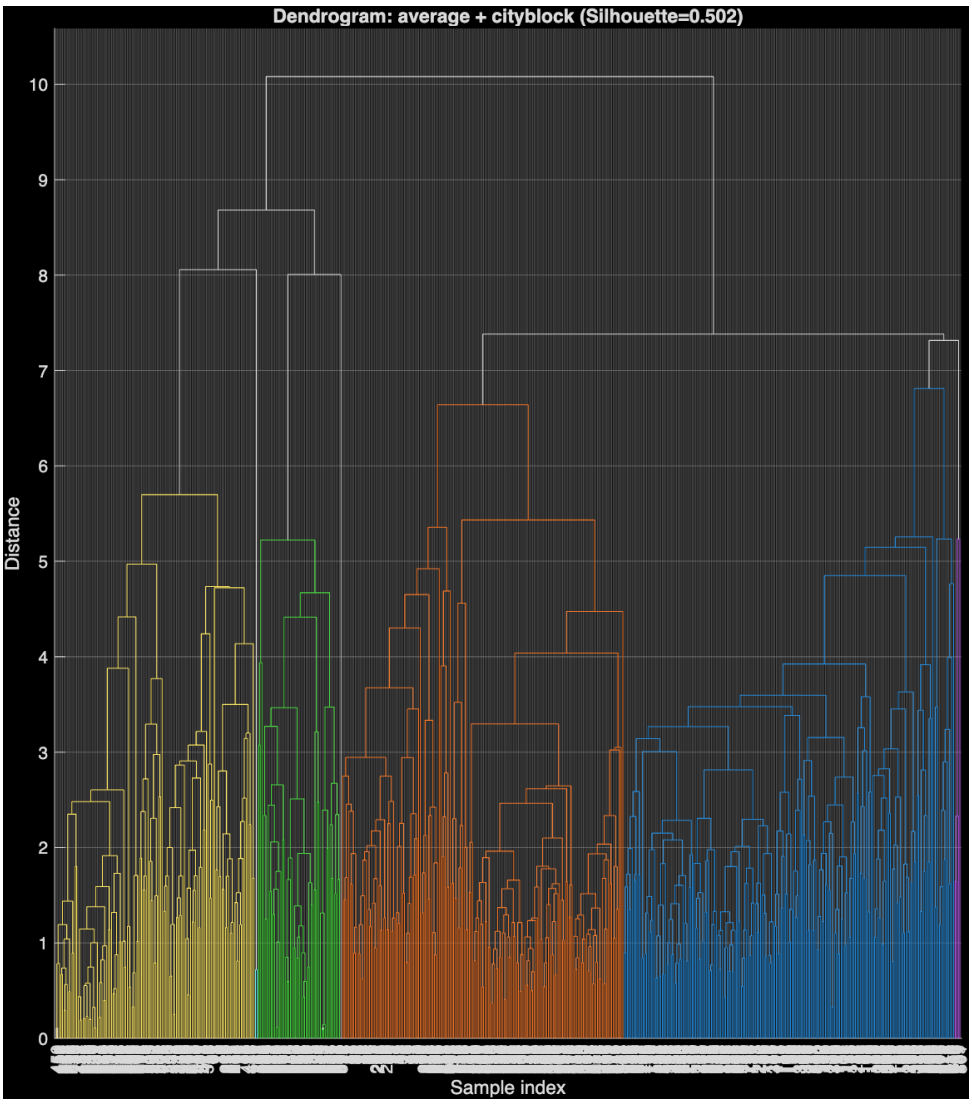


Figure 2.2: Dendrogramma del dataset Olive Oil tramite tecnica di clustering gerarchico Average.

In questo dendrogramma notiamo sempre un Silhouette score buono, ma notiamo proprio come metodi di raggruppamento gerarchico differenti producano risultati differenti sullo stesso insieme dei dati. Infatti in questo caso sono stati creati 4 cluster che identifichiamo mediante i colori: giallo, verde, arancione e blu. Possiamo dire che i cluster giallo e verde sono molto simili, mentre arancione e blu sono molto dissimili. Anche in questo caso il criterio di distanza utilizzato è stato quello della distanza CityBlock nel caso particolare in cui p=1.

2.2 Dendrogramma dei Campioni

Il seguente grafico mostra quanto sono correlate le varie feature del dataset andando a rappresentarle in un dendrogramma misurandone la similarità tramite la distanza di correlazione.

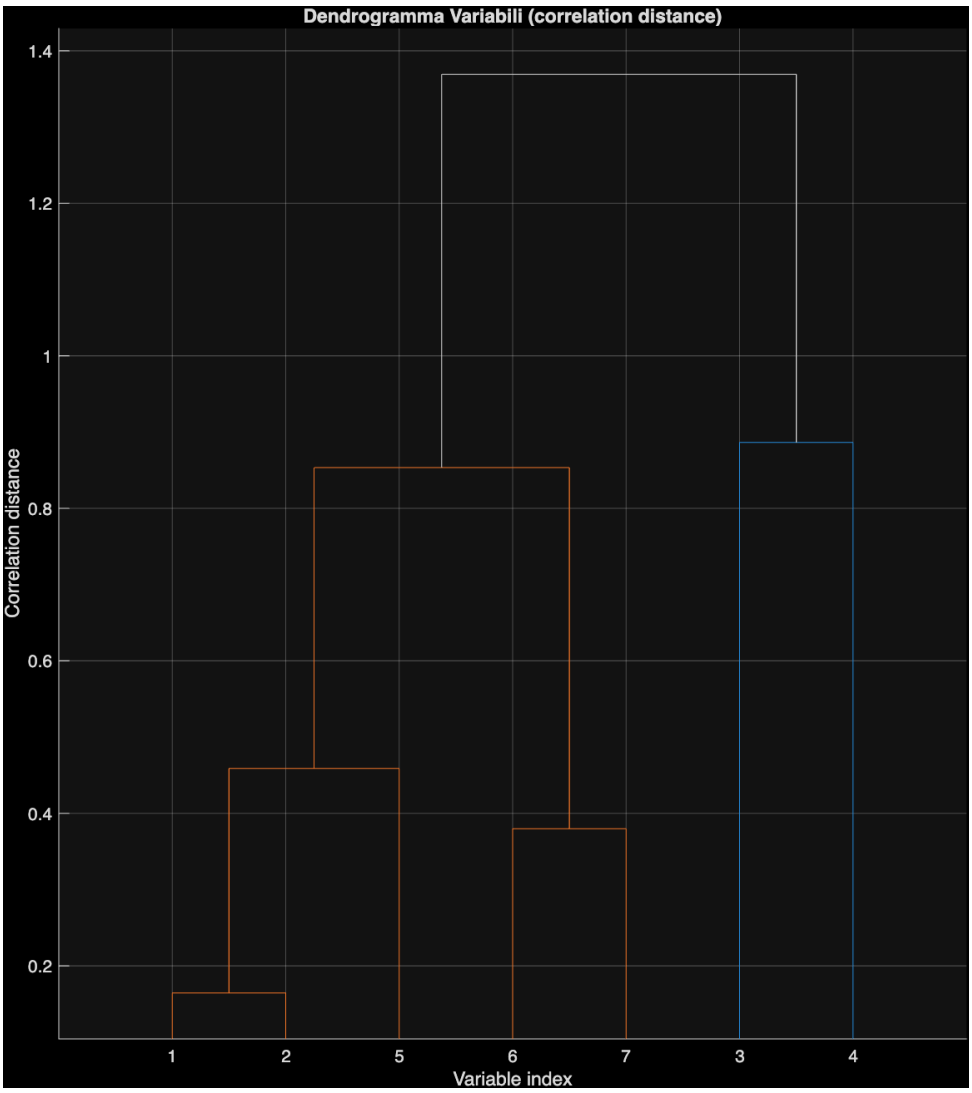


Figure 2.3: Dendrogramma delle variabili.

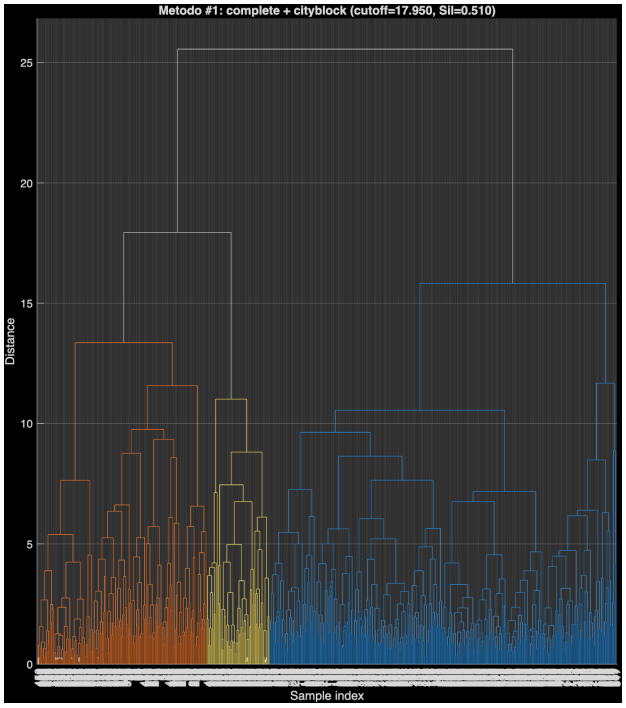
Come possiamo vedere, vi sono coppie di variabili molto simili quali: (1,2), (6,7), (3,4). E che le variabili (1,2,5,6,7) somigliano poco alle variabili (3,4).

2.3 Analisi PCA

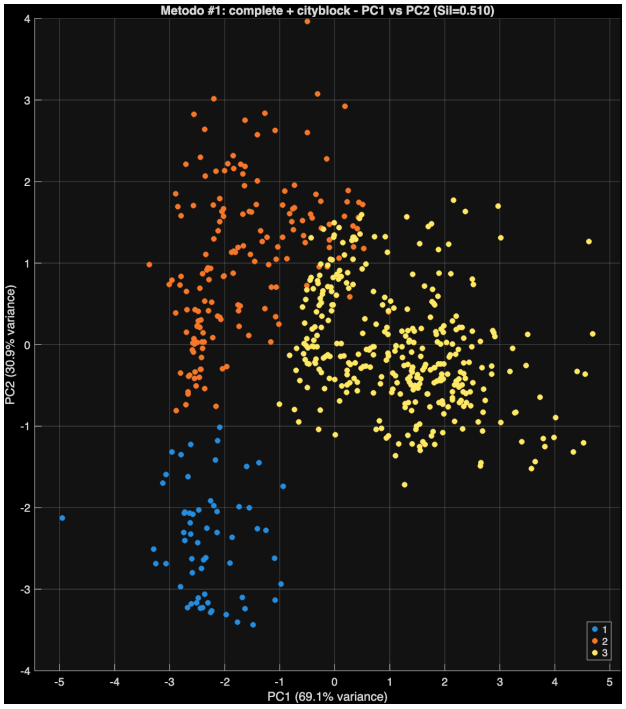
In questa fase andiamo a plottare gli scores nello spazio di PC1 e PC2 rendendo chiaro a che classe appartiene ogni campione.

2.3.1 Dendrogramma 1

É stato scelgo una soglia di cutoff pari a 17.950 e questo ci permette di individuare correttamente tre cluster che non verranno dunque fusi.



(a) Cut off

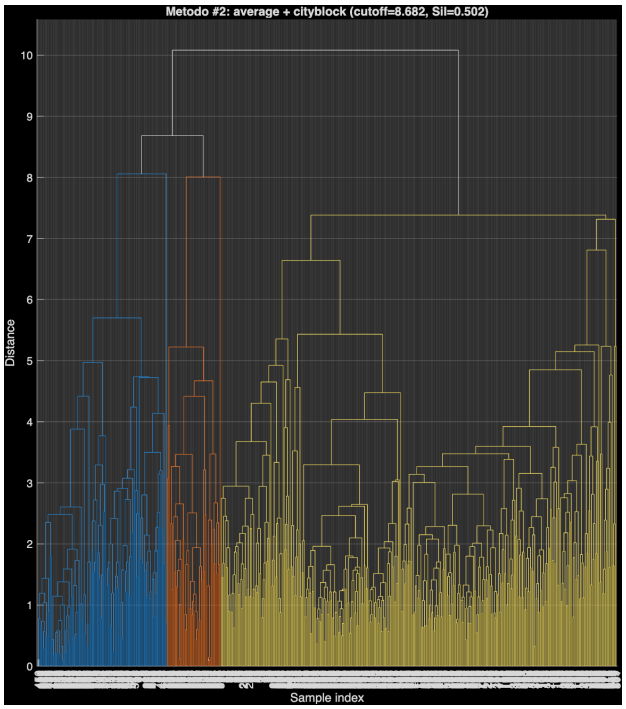


(b) Scores nello spazio PCA colorati per classe.

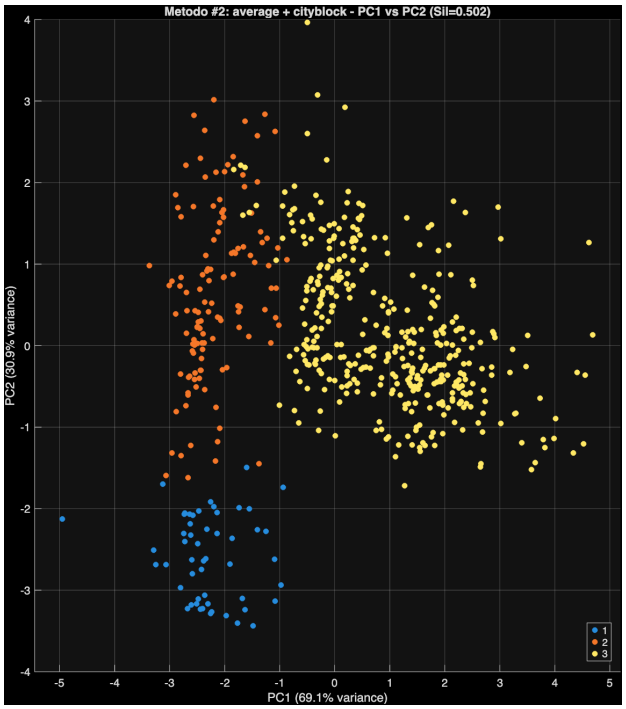
Nel grafico a destra invece vediamo la distribuzione dei campioni nello spazio delle componenti principali colorati per classi.

2.3.2 Dendrogramma 2

In questo caso invece la soglia del cutoff è stata scelta poco sopra 8.5 così da identificare correttamente 3 cluster ben separati.



(a) Cut off



(b) Scores nello spazio PCA colorati per classe.

Notiamo come anche la rappresentazione nello spazio delle componenti principali sia differente in quanto la tecnica di clustering utilizzata è differente.

2.4 Clustering su variabili

Questo grafico possiamo ottenerlo andando a trasporre la matrice dei dati e affiancando i dendrogrammi. Come è possibile notare dal grafico sull'asse y vi sono le variabili mentre sull'asse x vi sono i campioni.

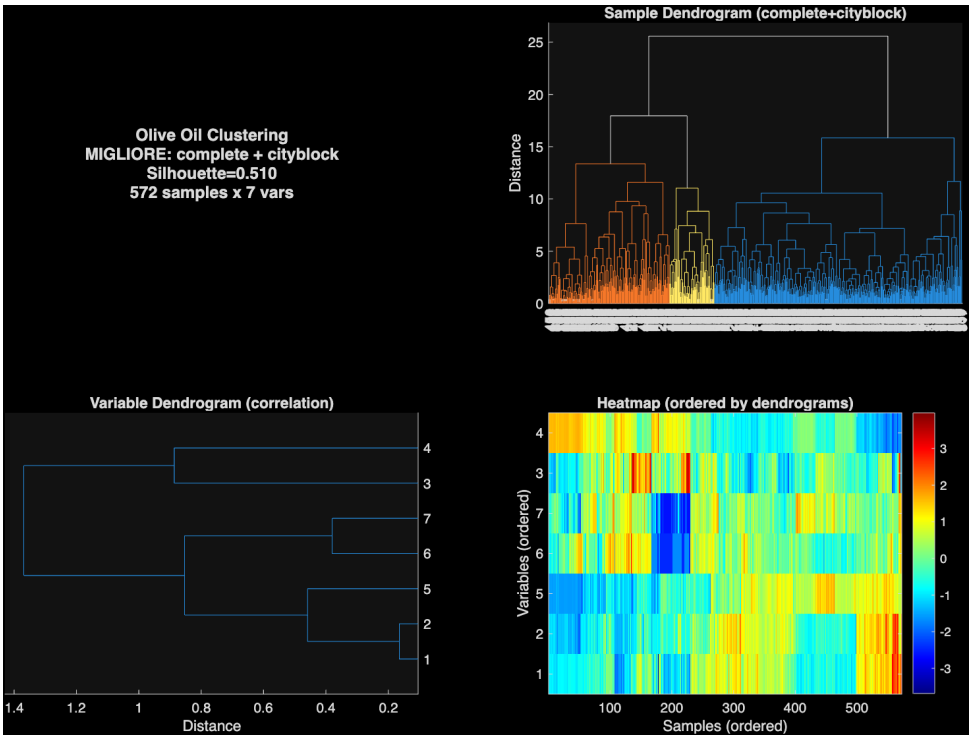


Figure 2.6: Dendrogramma delle variabili.

Osservando questo grafico siamo in grado di notare diversi pattern verticali, che indicano che certi gruppi di campioni condividono valori simili per quelle variabili come per esempio nella variabile 5 da 0 a 250 la variabile tende ad assumere valori tra -2 e -1 mentre tra 250 e 500 tende ad assumere valori tra 0 e 1. Possiamo anche dire che per la variabile 4 vi è un primo gruppo di campioni che ha il valore associato alto, poi i restanti basso. Per le variabili 6 e 7 vi è un ristretto gruppo di campioni che ha valori strettamente negativi, i restanti si aggirano intorno a 0 e 1. Per la altre variabili invece ci sono zone più discontinue nell'intervallo dei valori.

2.5 K-Means

K-Means è un algoritmo di apprendimento non supervisionato che non sa a priori il numero di cluster da creare, quindi è necessario specificarlo in input o farlo determinare tramite tecniche terze. Nel nostro caso è stato richiesto al K-Means di creare 8 cluster, tanti quante le categorie di olio.

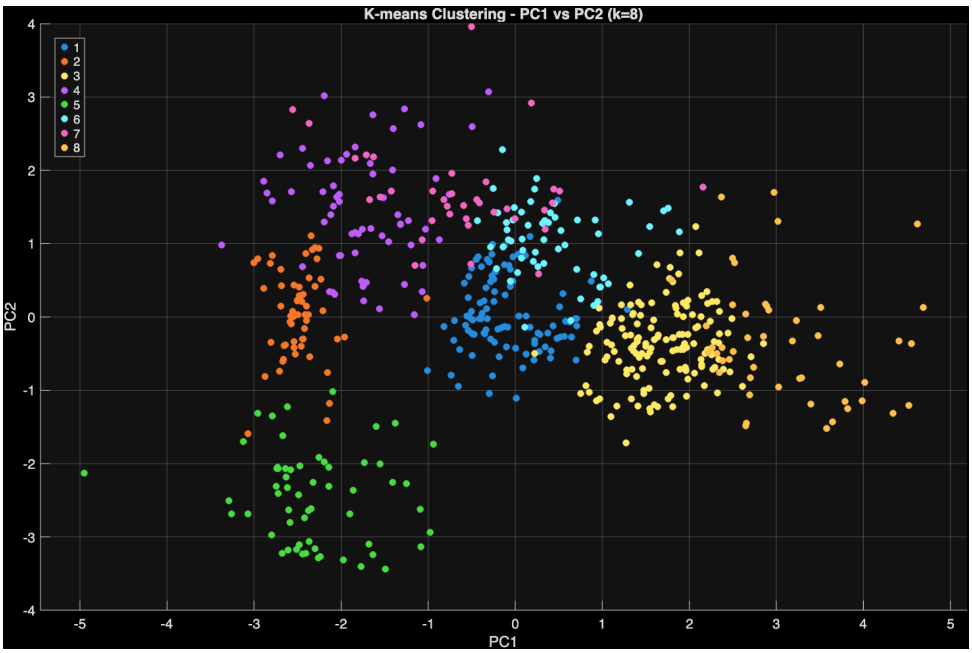


Figure 2.7: Clustering con K-Means.

Possiamo notare come questa tecnica di clustering abbia separato in maniera ancora differente i dati. Si riesce anche a notare la forma globulare tipica dei cluster di K-Means.
Ora vediamo il grafico di Silhouette di questo algoritmo.

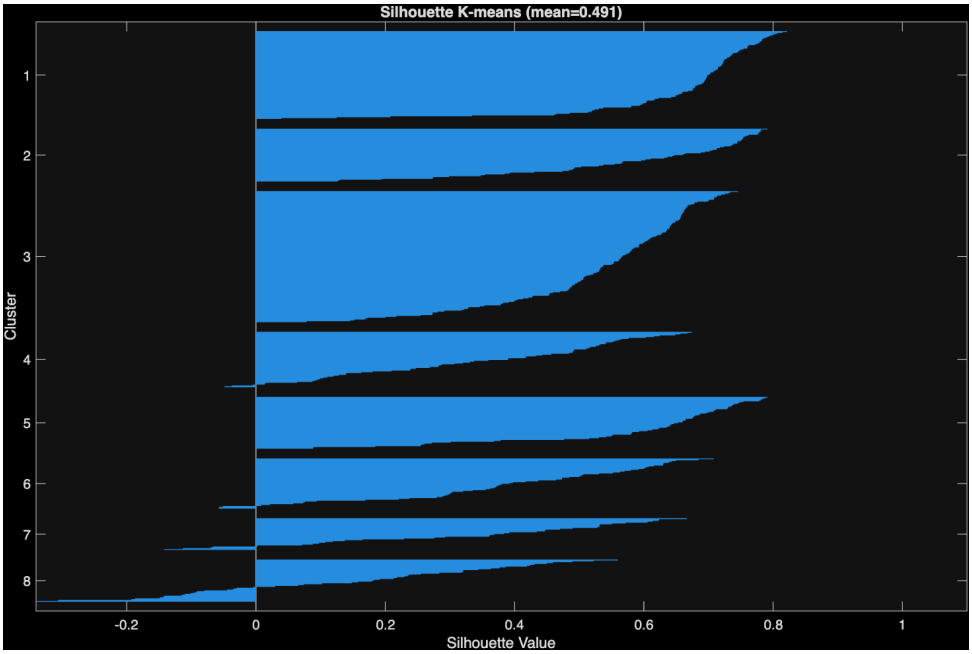


Figure 2.8: Silhouette K-Means

Lo score di silhouette complessivo è accettabile, il che indica una buona separazione dei dati.

2.6 DBSCAN

DBSCAN è invece un algoritmo che basa la creazione di cluster sull'analisi della densità nello spazio dei dati. Non ha bisogno di ricevere un parametro sul numero di cluster da creare in input in quanto lo determina da solo.

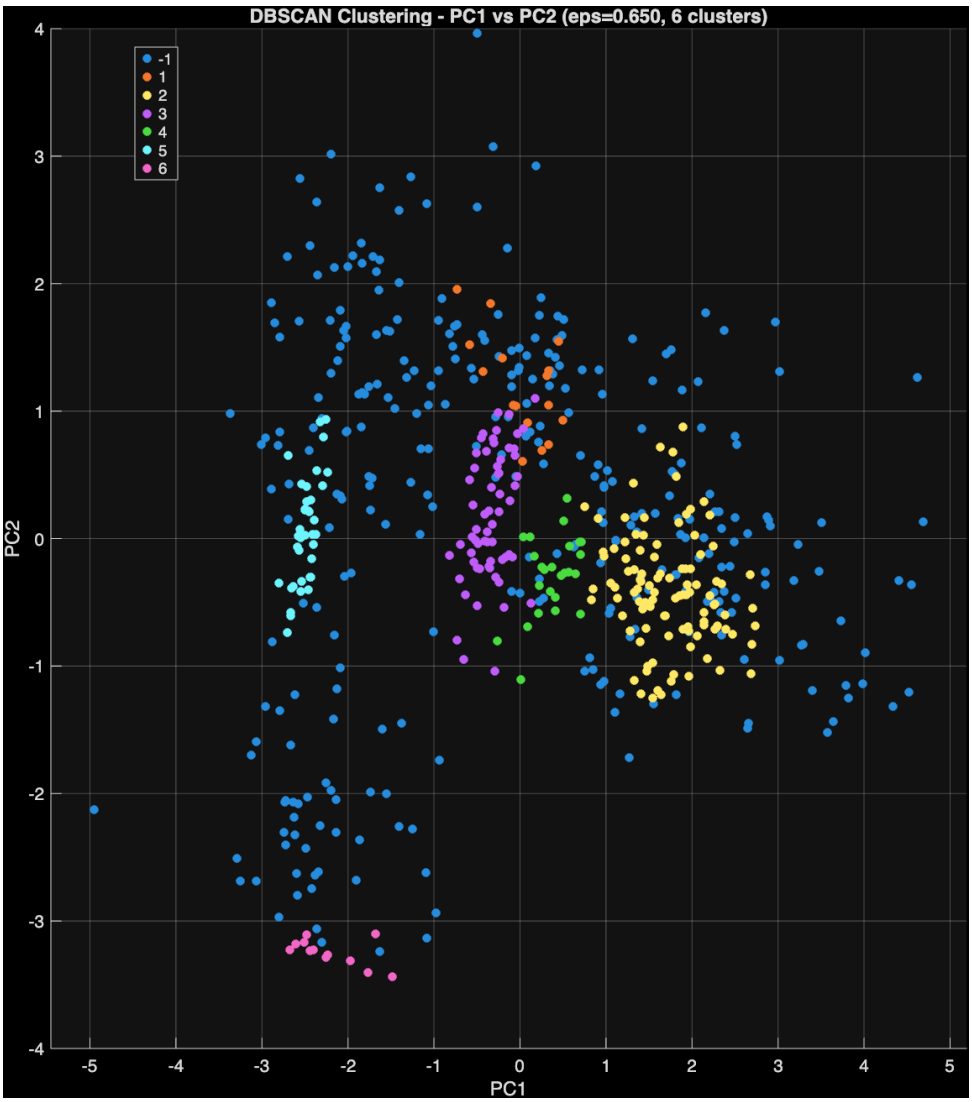


Figure 2.9: Clustering con DBSCAN

In questo grafico notiamo come vi siano cluster con forme molto meno globulari e che potrebbero, per assurdo, rispecchiare di più la reale distribuzione dei campioni nello spazio delle componenti principali dato che K-Means non lavora bene quando le classi hanno densità e numeri di campioni tanto differenti come avviene in questo dataset. Vengono dunque creati 6 cluster. Possiamo notare inoltre che, il parametro calcolato automaticamente dal programma come distanza di "ricerca" per DBSCAN è 0.650 e questo fa sì che la maggior parte dei sample vengano classificati come rumore andando a creare cluster solo dove vi sono aree veramente dense di punti. Probabilmente sarebbe meglio eseguire l'analisi con un valore di eps più alto.

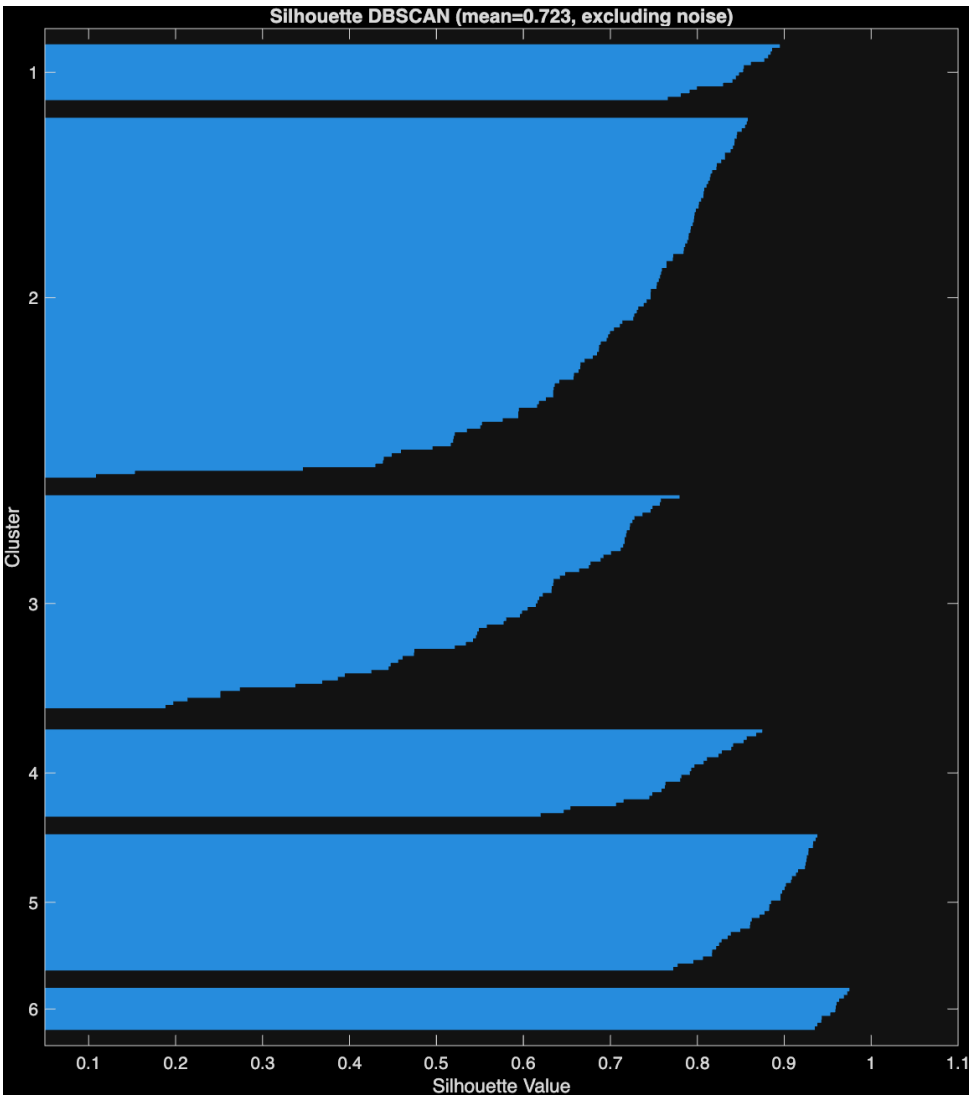


Figure 2.10: Silhouette DBSCAN

Inoltre dato che DBSCAN è un algoritmo di clustering che gestisce da solo il rumore, ovvero lo identifica e lo rimuove, non considerandolo nell'analisi del silhouette score notiamo avere un indice a valore quasi perfetto, in quanto valori estremamente vicini a uno potrebbero non essere ottimali e quindi circa 0.7 è perfetto.

2.7 Optics

Optics è un altro algoritmo di clustering basato sulla densità dei campioni nello spazio ed è considerata un'evoluzione di DBSCAN. A differenza di quest'ultimo va a considerare la raggiungibilità tra i punti nello spazio. Questo lo fa attraverso un parametro che è stato dato in input al nostro algoritmo a seguito di una divisione che ha restituito 23 come risultato.

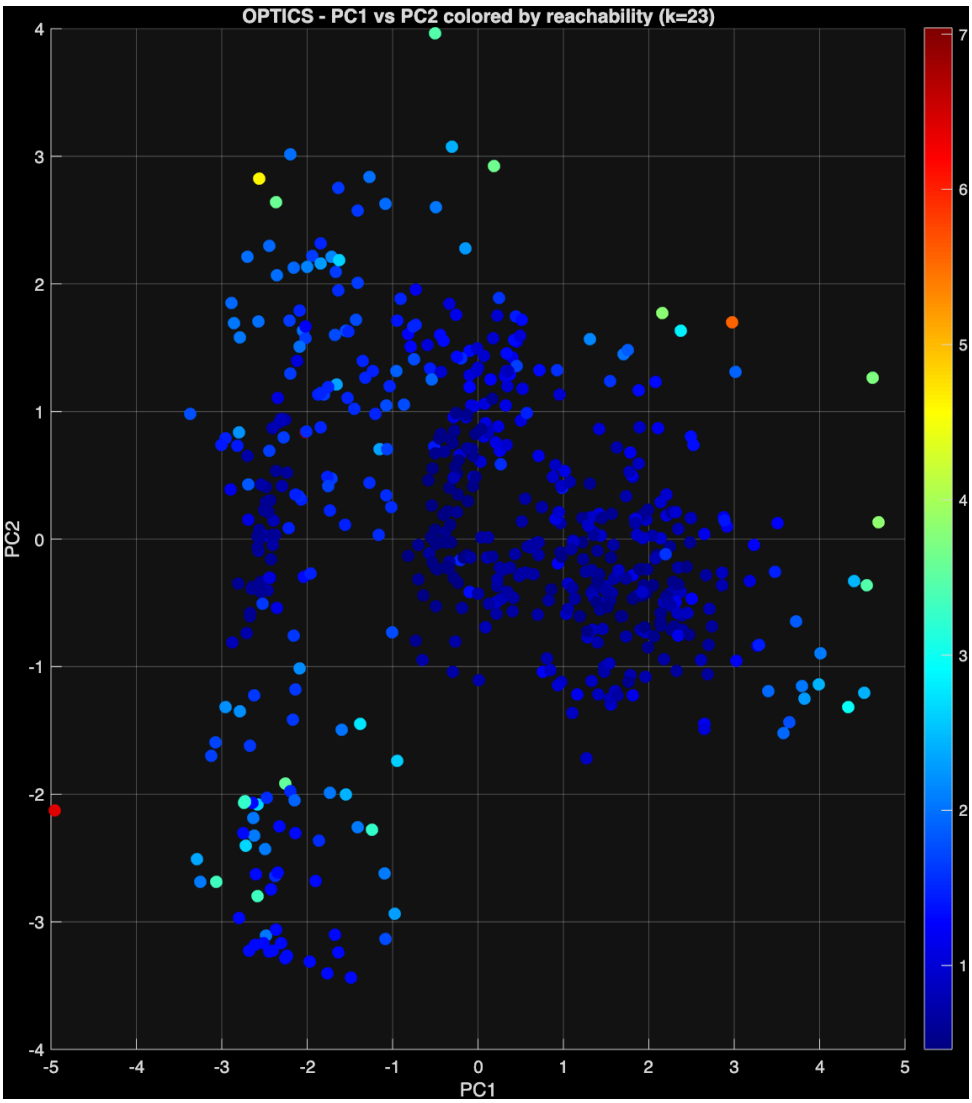


Figure 2.11: Silhouette con Optics

Questo grafico rappresenta i campioni colorati in base alla reachability distance. La scala cromatica indica la densità locale: i punti blu (valori bassi, 1-2) rappresentano campioni in regioni ad alta densità, facilmente raggiungibili dai loro vicini, mentre i punti verdi, gialli e rossi (valori >3) indicano campioni in regioni meno dense o isolati. La maggioranza dei campioni presenta valori di reachability bassi (blu), concentrati principalmente nella regione centrale del piano PC1-PC2, suggerendo la presenza di un nucleo denso di dati. I punti con reachability elevata (verde-giallo-rosso) sono dispersi prevalentemente nella periferia dello spazio ridotto, indicando potenziali outlier o campioni di transizione tra diverse strutture di densità.

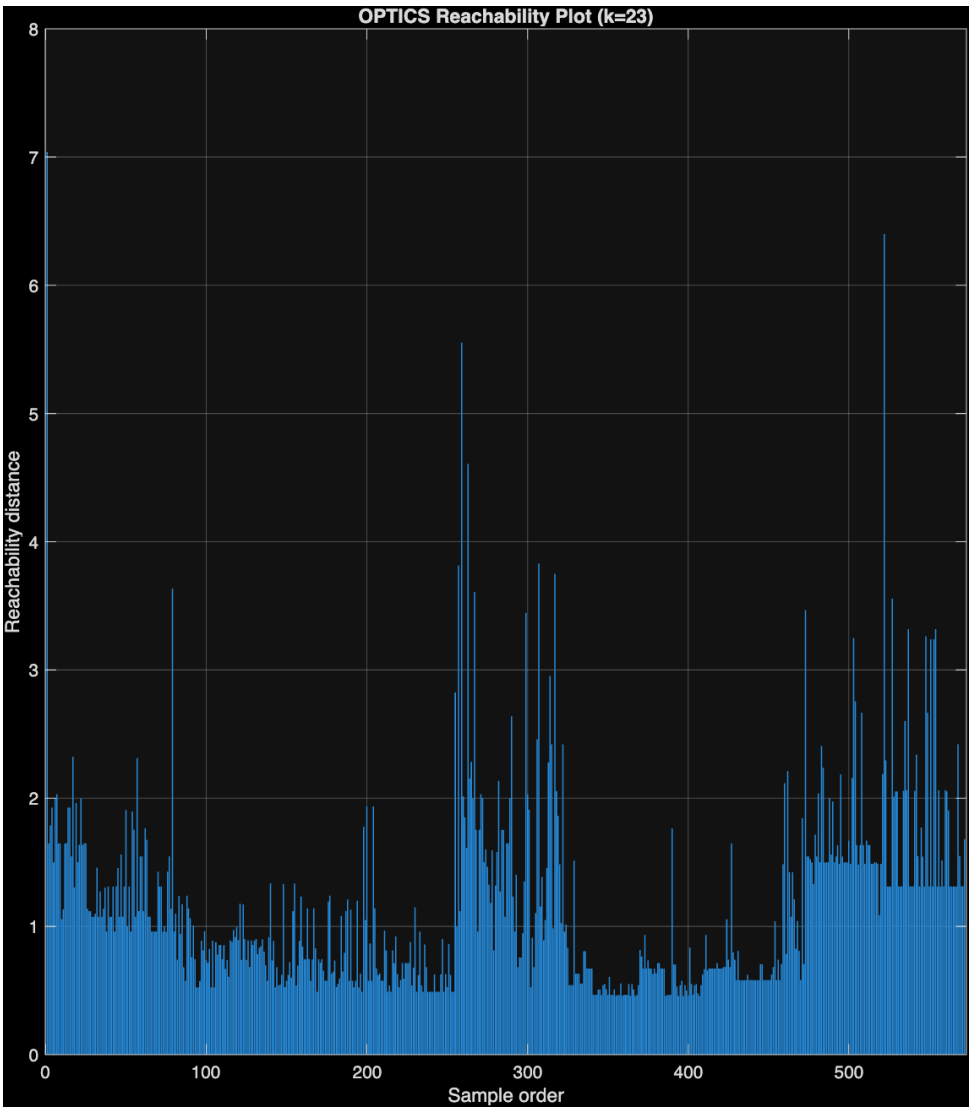


Figure 2.12: Reachability plot Optics

Questo grafico presenta il reachability plot caratteristico di OPTICS, che ordina i 572 campioni sull'asse x in base alla loro connettività, mentre l'asse y riporta la reachability distance di ciascun campione. Questo grafico permette di identificare visivamente la struttura gerarchica dei cluster presenti nei dati. Le valli (regioni con valori bassi di reachability) indicano cluster densi:
Campioni 0-100: cluster con densità moderata (reachability 1-2) Campioni 250-350: cluster molto denso con reachability prevalentemente <1 Campioni 450-550: cluster con densità variabile
I picchi (valori elevati di reachability) rappresentano separazioni tra cluster o punti outlier. Si osservano picchi particolarmente pronunciati intorno ai campioni 250-300 (reachability 5.5) e 500 (reachability 6.5), indicando campioni fortemente isolati o punti di transizione tra strutture di densità diverse.

3

Conclusioni

In conclusione, sono stati riportati tanti risultati, ma in questa sezione andiamo ad effettuare un confronto delle tecniche di clustering impiegate nell'analisi.
Come primo confronto possiamo prendere l'indice di silhouette e basandoci su questo possiamo dire che nelle varie configurazioni di clustering gerarchico solitamente questo valore era piuttosto basso, anche sotto 0.3 a volte. Questo perché questa tipologia di clustering è molto sensibile agli outliers, che come abbiamo visto tramite anche optics sono presenti all'interno del dataset. Per quanto riguarda K-Means invece si è ottenuto un silhouette score perfetto, anche se pure questo algoritmo soffre di outliers e tende a creare cluster globulari che, in caso di densità variabile nei dati del dataset, comporta la creazione di cluster che non rispecchiano la reale distribuzione dei dati. Questo infatti avviene. DBSCAN è invece l'algoritmo che ha riportato il silhouette score maggiore rispetto agli altri andando a creare 6 cluster compatti.
Un'altra osservazione che possiamo fare che sfavorirebbe l'utilizzo di tecniche di clustering gerarchico è che questi non hanno una complessità computazionale scalabile in contesti di dataset grandi mentre DBSCAN e K-Means scalano meglio.
Hanno però il pro che non è necessario nessun parametro, scegliamo solamente la combinazione che vogliamo generare, senza dover specificare il numero di cluster da creare o l'eps o il minPts.
In generale, la scelta della miglior tecnica di clustering dipende sempre dal dataset che stiamo analizzando e dall'obiettivo che ci poniamo. Sicuramente un algoritmo di clustering gerarchico possiamo impiegarla quando non abbiamo un numero elevato di campioni o quando non sappiamo la struttura dei dati del nostro dataset e vogliamo comprenderla.
Un clustering di tipo partitioning come K-Means invece possiamo usarlo quando sappiamo come assegnare correttamente il valore di k, quindi sappiamo quante classi ci sono e quando i cluster sono globulari o sferici, ben separati e non vi sono troppi outlier.
Mentre una tecnica di clustering basata su densità possiamo impiegarla quando non sappiamo il numero di cluster da creare, oppure quando è presente un gran rumore sui dati e quando i cluster possono assumere forme qualsiasi.